

Introduction to Big Data Assignment-2

Name: Faizan Mulla

Roll No: 21F1003885

Problem Statement

The goal of this project was to develop an automated solution for counting the lines of text files uploaded to a Google Cloud Storage (GCS) bucket. By utilizing Google Cloud Functions, the solution leverages serverless architecture to handle file uploads efficiently and scalably. This report outlines the steps taken to set up the project, enable necessary APIs, write and deploy the Python code, and test the functionality.

Solution

The solution involves setting up a Google Cloud Function that triggers when a new text file is uploaded to a designated GCS bucket.

The function reads the file, counts the number of lines, and stores the result in a new file within the same bucket. This approach automates the line-counting process, making it suitable for large-scale applications where multiple files need processing without manual intervention.

Implementation Details

1. GCP Environment Setup

- **Create GCP Project:** Accessed Google Cloud Console and created a **new** project named **"IBD-GA3"**. Enabled necessary APIs including *Cloud Functions API*, *Cloud Build API*, and *Cloud Storage API*.
- **Cloud Storage Setup:** Created a GCS bucket named **"ibd-ga3-bucket"** in the **"asia-south1"** region, configured with the standard storage class.

2. Cloud Function Configuration

- Create Cloud Function: Navigated to Cloud Functions in the GCP Console, created a new function with the following details:
 - **Function Name:** "count-lines"
 - **Region:** asia-south1
 - **Trigger Type:** Cloud Storage
 - **Event Type:** Finalize/Create
 - **Bucket:** Selected "ibd-ga3-bucket"
 - Set the runtime environment to **Python 3.10** and define the entry point as ``count_lines``.

3. Write Function Code

- Developed the following Python code to count the lines in a text file:

```
from google.cloud import storage

def count_lines(event, context):
    """Triggered by a change to a Cloud Storage bucket."""
    file_name = event["name"]
    bucket_name = event["bucket"]

    # Ignore files that start with "results_"
    if file_name.startswith("results_"):
        print(f"Ignoring file {file_name} as it starts with 'results_'")
        return

    # Initialize client
    storage_client = storage.Client()

    # Get bucket and file
    bucket = storage_client.bucket(bucket_name)
    blob = bucket.get_blob(file_name)

    # Read content and count lines
    content = blob.download_as_text()
    line_count = len(content.splitlines())

    # Create result message
    result = f"File {file_name} contains {line_count} lines"
    print(result) # This will appear in logs

    # Save result to new file
    result_blob = bucket.blob(f"results_{file_name}")
    result_blob.upload_from_string(result)
```

4. Create `requirements.txt`

- Added the required library to interact with GCS: `"google-cloud-storage>=2.0.0"`

Execution Process

1. Deploy Cloud Function

- Clicked "Deploy" in the Cloud Functions console.
- Waited for the function to deploy successfully, which took a few minutes.

2. Test the Function

- Uploaded a text file (`input.txt`) to the "ibd-ga3-bucket".
- Waited for the function to process the file automatically.
- Observed the creation of a new file `results_input.txt` in the bucket containing the line count result.
- Verified the output by downloading `results_input.txt`, which displayed the correct line count of the uploaded file.

Results

The function successfully counts the lines of any uploaded text file in the specified bucket and stores the results in a new file **prefixed with "results_"**.

We can now just download it and add it to the submission folder.

Example Result:

For a file named `input.txt` with 9 lines, the output stored in `results_input.txt` would be:

...

File example.txt contains 9 lines

...

Relevant Screenshots

1. Cloud Function Configuration:

ga2-cloud-lines

Cloud Run function (1st gen)

Version 2, deployed at Oct 23, 2024, 6:55:30 P...

METRICS

DETAILS

SOURCE

VARIABLES

TRIGGER

PERMISSIONS

LOGS

TESTING

Runtime: Python 3.10

Entry point: count_lines

EDIT

main.py

requirements.txt

```
1 from google.cloud import storage
2
3 def count_lines(event, context):
4     """Triggered by a change to a Cloud Storage bucket."""
5     file_name = event['name']
6     bucket_name = event['bucket']
7
8     # Ignore files that start with "results_"
9     if file_name.startswith("results_"):
10         print(f"Ignoring file {file_name} as it starts with 'results_'")
11         return
12
13     # Initialize client
14     storage_client = storage.Client()
15
16     # Get bucket and file
17     bucket = storage_client.bucket(bucket_name)
18     blob = bucket.get_blob(file_name)
19
20     # Read content and count lines
21     content = blob.download_as_text()
22     line_count = len(content.splitlines())
23
24     # Create result message
25     result = f"File {file_name} contains {line_count} lines"
26     print(result) # This will appear in logs
27
28     # Save result to new file
29     result_blob = bucket.blob(f"results_{file_name}")
30     result_blob.upload_from_string(result)
31
32     return result
```

ga2-cloud-lines

Cloud Run function (1st gen)

Version 2, deployed at Oct 23, 2024, 6:55:30 P...

METRICS

DETAILS

SOURCE

VARIABLES

TRIGGER

PERMISSIONS

LOGS

TESTING

Runtime

Python 3.10

Entry point *

count_lines

SAVE AND REDEPLOY

CANCEL

TEST FUNCTION

main.py

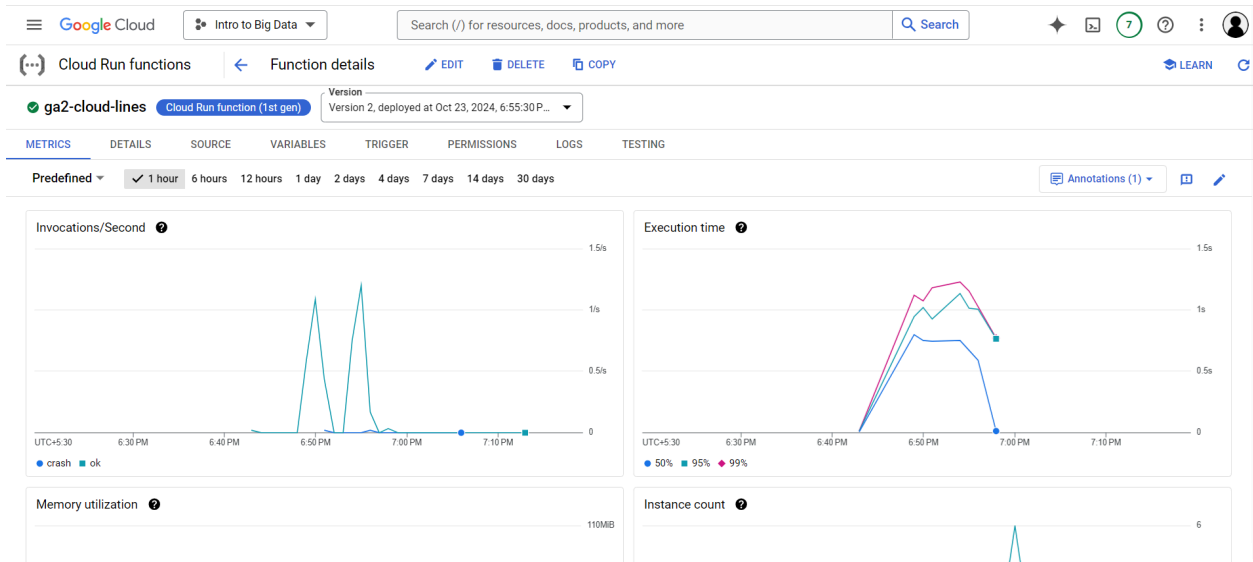
requirements.txt

Press Alt+F1 for Accessibility Options.

1 Functions-framework==3.*

2 google-cloud-storage

3



2. Storage Bucket Setup:

Bucket details GO TO PATH REFRESH LEARN

ibd-ga2-bucket

Location: asia (multiple regions in Asia) Storage class: Standard Public access: Not public Protection: Soft Delete

OBJECTS CONFIGURATION PERMISSIONS PROTECTION LIFECYCLE OBSERVABILITY INVENTORY REPORTS OPERATIONS

Folder browser

ibid-ga2-bucket

CREATE FOLDER UPLOAD TRANSFER DATA OTHER SERVICES

Filter by name prefix only Filter objects and folders Show Live objects only

Name	Size	Type	Created	Storage class	Last modified
input.txt	373 B	text/plain	Oct 23, 2024, 6:57:26 PM	Standard	Oct 23, 2024, 6:57:26 PM

3. Function Execution Logs:

ga2-cloud-lines

Cloud Run function (1st gen)

Version
Version 2, deployed at Oct 23, 2024, 6:55:30 P...

METRICS	DETAILS	SOURCE	VARIABLES	TRIGGER	PERMISSIONS	LOGS	TESTING
<div>Logs</div> <div>Severity: Default</div> <div>Filter Search all fields and values</div>							
SEVERITY	TIMESTAMP	SUMMARY					
>	2024-10-23 18:55:01.845 IST	ga2-cloud-lines wc5mu4u8mqbf "domain": "global",					
>	2024-10-23 18:55:01.845 IST	ga2-cloud-lines wc5mu4u8mqbf "reason": "invalid"					
>	2024-10-23 18:55:01.845 IST	ga2-cloud-lines wc5mu4u8mqbf }					
>	2024-10-23 18:55:01.845 IST	ga2-cloud-lines wc5mu4u8mqbf]					
>	2024-10-23 18:55:01.845 IST	ga2-cloud-lines wc5mu4u8mqbf }					
>	2024-10-23 18:55:01.845 IST	ga2-cloud-lines wc5mu4u8mqbf }					
>	2024-10-23 18:55:01.845 IST	ga2-cloud-lines wc5mu4u8mqbf }					
>	2024-10-23 18:55:01.845 IST	ga2-cloud-lines wc5mu4u8mqbf : ('Request failed with status code', 400, 'Expected one of', <HTTPStatus.OK: 200>)					
>	2024-10-23 18:55:01.846 IST	ga2-cloud-lines wc5mu4u8mqbf Function execution took 657 ms, finished with status: 'crash'					
>	2024-10-23 18:55:30.388 IST	Cloud Functions UpdateFunction asia-south1:ga2-cloud-lines faizanamulla69@gmail.com (@type: type.googleapis.com/google.cloud.audit.AuditLog, authenticationInfo: {...}, methodName: google...					
>	2024-10-23 18:57:28.005 IST	ga2-cloud-lines lnh8ji684pnj Function execution started					
>	2024-10-23 18:57:28.512 IST	ga2-cloud-lines lnh8ji684pnj File input.txt contains 9 lines					
>	2024-10-23 18:57:28.766 IST	ga2-cloud-lines lnh8ji684pnj Function execution took 761 ms, finished with status: 'ok'					
>	2024-10-23 18:57:28.781 IST	ga2-cloud-lines lnh8kpx3vehl Function execution started					
>	2024-10-23 18:57:28.784 IST	ga2-cloud-lines lnh8kpx3vehl Ignoring file results_input.txt as it starts with 'results_'					
>	2024-10-23 18:57:28.785 IST	ga2-cloud-lines lnh8kpx3vehl Function execution took 3 ms, finished with status: 'ok'					
No newer entries found matching current filter.							

4. Generated Output File:

ibid-ga2-bucket

Location: asia (multiple regions in Asia)

Storage class: Standard

Public access: Not public

Protection: Soft Delete

OBJECTS	CONFIGURATION	PERMISSIONS	PROTECTION	LIFECYCLE	OBSERVABILITY	INVENTORY REPORTS	OPERATIONS																								
<div>Folder browser</div> <div>ibid-ga2-bucket</div> <div>CREATE FOLDER UPLOAD TRANSFER DATA OTHER SERVICES</div> <div>Filter by name prefix only Filter objects and folders Show Live objects only</div> <table><thead><tr><th></th><th>Name</th><th>Size</th><th>Type</th><th>Created</th><th>Storage class</th><th>Last modified</th><th></th></tr></thead><tbody><tr><td><input type="checkbox"/></td><td>input.txt</td><td>373 B</td><td>text/plain</td><td>Oct 23, 2024, 6:57:26 PM</td><td>Standard</td><td>Oct 23, 2024, 6:57:26 PM</td><td>⬇ ⋮</td></tr><tr><td><input type="checkbox"/></td><td>results_input.txt</td><td>31 B</td><td>text/plain</td><td>Oct 23, 2024, 6:57:28 PM</td><td>Standard</td><td>Oct 23, 2024, 6:57:28 PM</td><td>⬇ ⋮</td></tr></tbody></table>									Name	Size	Type	Created	Storage class	Last modified		<input type="checkbox"/>	input.txt	373 B	text/plain	Oct 23, 2024, 6:57:26 PM	Standard	Oct 23, 2024, 6:57:26 PM	⬇ ⋮	<input type="checkbox"/>	results_input.txt	31 B	text/plain	Oct 23, 2024, 6:57:28 PM	Standard	Oct 23, 2024, 6:57:28 PM	⬇ ⋮
	Name	Size	Type	Created	Storage class	Last modified																									
<input type="checkbox"/>	input.txt	373 B	text/plain	Oct 23, 2024, 6:57:26 PM	Standard	Oct 23, 2024, 6:57:26 PM	⬇ ⋮																								
<input type="checkbox"/>	results_input.txt	31 B	text/plain	Oct 23, 2024, 6:57:28 PM	Standard	Oct 23, 2024, 6:57:28 PM	⬇ ⋮																								

```
results_input.txt X
1 File input.txt contains 9 lines
```